Using Proper Names to Cluster Documents

Dan Winchester & Mark Lee
School of Computer Science
University of Birmingham
B15 2TT
England
{dyw/mgl@cs.bham.ac.uk}

Abstract

Proper Names are a frequent occurrence in all types of natural language text. However, the treatment of proper names is an area under-researched by Natural Language Processing. One particular problem is how to link information about the same entity referred to by possibly different proper names in several documents. In this paper we describe a prototype system which first pre-processes individual documents using a simple name-conflation algorithm and then uses an adaptation of Schutze's context-group discrimination algorithm to cluster documents that are judged to contain references to the same named entity. We use this system to assess the potential utility of different contextual cues to the task.

Introduction

Proper Names occur frequently in all types of natural language text and their comprehension is central to text understanding. This is particularly true in the news text genre where it has been estimated that proper names comprise around 10% of the text (Coates-Stephens, 1992). However, unlike other linguistic categories, proper names are poorly represented in lexical resources and the area remains under-researched. One specific problem in this area is the resolution of proper name reference across a collection of documents, a problem that has been referred to as cross-document coreference (Bagga & Baldwin, 1998a; Ravin & Kazi, 1998). Our work is primarily concerned with developing practical computational solutions to this problem. Such a step is necessary to link information about the same entity referred to by possibly different proper names in several documents and conversely, to distinguish between different entities that share the same, or a similar name. Cross-document coreference is a problem that will become increasingly apparent as NLP technology is applied to ever-larger collections of documents since, as collection size and generality increase, so does the potential for ambiguity.

As human language users, we are seldom aware of ambiguity in the intended referent of a name, just as we are seldom aware of ambiguity in the intended *sense* of a common noun, verb, etc. Context provides the information we need to identify the intended referent. However, isolate a name entirely from context (in its broadest sense) and the ambiguity becomes apparent. It is then impossible to offer anything but a default choice as to the correct referent of the name. This is because names do not, in fact, possess the unique reference that we typically ascribe to them. Rather, there is a many to many mapping between names and entities; many people share the name "*John Smith*" and there are a number of predictable variant forms of the name that can be used for any particular John Smith *('J. Smith', 'Smith', 'John'*, etc). Within a single discourse, such ambiguity is seldom problematic. Identically or similarly named entities seldom appear in the same context, and, when they do, competent language users will generally distinguish between them explicitly. In effect this allows an assumption of *one referent per discourse*. Across a collection of documents the same safe assumptions cannot be made. Without examining the context of each occurrence, it is impossible to enumerate a priori the number of entities that share a particular proper name or to determine which occurrences of the name refer to which entity. Any task that aims to step outside of the confines of individual documents and link information about an entity from different sources must acknowledge and overcome this ambiguity.

Potential Applications

Thus the potential applications for cross-document coreference techniques are those that require names to be linked across a corpus, not on the basis of orthographic similarity, but because of their intended referent. One obvious application is in the broad field of Information Retrieval (IR). Reliable cross-document coreference techniques would allow references to the same entity to be linked across a corpus. This could be used to improve indexing, organise search results or to offer the user links from one document containing a particular entity to others containing the same, and not just a similarly named, entity.

Significantly, there are also indications that the field Information Extraction is beginning to consider the broad problem of tracking information (particularly that regarding entities) across document boundaries. The Automatic Context Extraction (ACE) initiative includes an Entity Detection and Tracking (EDT) component and there are very clear indications that the eventual aim is to collate information about entities from multiple sources (ACE 1999).

Objectives

There are two primary objectives for this research. The first is to develop and evaluate a system that addresses the cross-document coreference problem on a reasonably large scale. The second is to use this system to assess the potential utility of different contextual cues to the task. The treatments of cross-document coreference that have been offered to date are either limited by the scale of their investigation (Bagga and Baldwin, 1998a) or by the complexity of the processing involved (Coates-Stephens, 1992). What is needed is a fuller investigation of the domain based on a method that is robust and requires little or no deep processing. We present here a method that is intended to meet these criteria. The current system relies on second-order *proper name* co-occurrence statistics to build representations of the context in which an ambiguous proper name occurs. A clustering algorithm then merges similar occurrences so that each cluster represents a different entity. The choice of second-order proper-name co-occurrence as a representational basis arises from two premises: 1) That second-order co-occurrence is more robust than first-order co-occurrence; and 2) that an entity is more reliably identifiable by the other named entities that typically surround it than by less unique contextual features (open class words). We have tested these premises by using the general framework described below to compare various possible representations of context.

Thus, the problem of cross-document coreference is one that has received little attention, but it is nevertheless one to which a solution should prove valuable. The challenge is to develop techniques that can remove ambiguity from names by sorting occurrences of a proper name and its accepted variants throughout a corpus into a number of groups, each containing references to a single distinct entity. The following section provides details of the method used to achieve this.

Method

All of our experiments used a small corpus of 6000 Wall Street Journal articles (2.5 million Words). These documents were first processed using a Named Entity Recognition tool developed by Sharp Laboratories of Europe. Named Entities were assigned one of three SGML tags categorising them as either person (PER), location (LOC) or organisation (ORG). Our research is intended for use in large corpora since it is likely that the size of a corpus is proportional to the amount of ambiguity in names. The 2.5 million word WSJ corpus, however, is relatively small. In order to simulate the ambiguity of a larger corpus it was necessary to create pseudo-ambiguity in the data. This was done very simply by creating namesets - sets of names that shared common features. Each nameset consisted of all those documents containing a proper name that matched a certain regular expression. For example, in the case of PER proper names the criteria was the presence of a certain surname, as determined by a regular expression such as '* Smith'. This method was used to delineate 30 namesets; 10 of each entity type (PER, LOC, ORG) and an answer key was produced for each by manual annotation. A certain amount of intervention was necessary in the selection of these namesets to ensure that they varied on a number of criteria. Specifically, we aimed to define namesets that varied in size (number of occurrences), diversity (number of distinct entities) and distribution (how occurrences were split between the possible entities).

Pre-Processing

Before we address the cross-document coreference task 'proper', it is possible to greatly simplify the problem via a pre-processing stage that partially addresses *within-document* coreference by linking together multiple occurrences of the same proper name and its variants within a single document. This, in conjunction with the assumption of one referent per discourse above, effectively reduces the problem from the level of individual occurrences of a proper name to the document level. We are now trying to cluster *documents* that contain references to the same entity.

In this stage, each document in the corpus is processed in turn. Where appropriate, variant name forms are conflated and simple information (such as gender cues) associated with a name is stored. Conflation of variant names is achieved via three heuristics that utilise the regularities present in variant forms of proper names of the three different entity types. This pre-processing stage has three motivations: it simplifies the co-occurrence statistics; it links the different contexts of occurrence within a document; and it associates reduced forms of a name with the most canonical form used in that document. Once this conflation stage is complete, indexes can be built from which co-occurrence statistics are derived in the second phase.

Context Group Discrimination

Thus, the task is to divide occurrences (where an occurrence now comprises all the individual instances of a name in a single document) of an ambiguous name into different classes, with each class representing one of the possible referents of the name. This task shares a number of features with the problem of Word Sense Disambiguation (WSD). However, while it is possible in WSD to enumerate *a priori,* the possible senses of an ambiguous word, outside of small domain-specific collections it is impossible to enumerate in advance the referents that share a particular name. This fact imposes certain limitations on the types of technique that can be employed for cross-document coreference. To satisfy these constraints, we use an adaptation of context-group discrimination, a technique first employed by Schutze (1998) for the task of 'Automatic Word Sense *Discrimination'*. This is a WSD method designed to group occurrences of an ambiguous word into a number of classes (representing different senses) on the basis of contextual similarity. Crucially, however, the method does not attempt to label the resultant classes from a predefined list of senses. Thus, the overall framework of the method is ideal to group contextually similar occurrences of a name into an unknown number of unlabelled classes. We have adapted this approach to the domain of proper names and, by varying the nature of the representations used, attempted to determine which contextual features would prove most useful to the disambiguation of names.

The method can be classed broadly as a vector space model based on second order co-occurrence statistics. There are three principal stages that cumulatively generate representations of: a) individual terms in the immediate context of the occurrence, `*term vectors*'; b) the entire context in which a name occurs, `*context vectors*'; and c) clusters of similar contexts that represent different entities referred to by a similar name, `*entity vectors*'. All of these representations exist in the same high-dimensional, real-valued vector space. Each dimension of vector space is a term (a word or name) that occurs elsewhere in the corpus. For example, in the final formulation of the model, the dimensions of vector space are all the proper names (within certain frequency boundaries) that occur throughout the corpus. For simplicity, it is this formulation that we will describe in the remainder of this section. We have, however, used the same general framework to evaluate representations based on first-order co-occurrence and non-name co-occurrence (i.e. *words* as dimensions of Vector Space) and these results will be referred to briefly below.

First, for each occurrence of the proper name of interest a vector representation is built in two stages. Term Vectors are created for each other proper name that occurs within a certain distance of the name of interest. A term vector stores co-occurrence statistics for that proper name across the whole corpus. So, the entry for name *x* in the vector for name *y* contains the number of times that *x* and *y* co-occur throughout the corpus. Thus names that typically appear in the same 'company' should reflect this similarity in the vectors that represent them. Context Vectors are then created to represent the entire context in which the target name occurs. This is done by summing the term vectors of all those proper names that co-occur directly with the ambiguous name. Before summing, individual term vectors are weighted with the standard inverse document frequency algorithm (Salton & McGill, 1983). After summing, the resulting centroid averages the direction of the set of term vectors, consolidating second-order co-occurrence information about this occurrence of the ambiguous name.

Once a context vector has been built for each occurrence of the name of interest, we can attempt to cluster similar context vectors to produce representations of the entities that these different occurrences refer to – Entity Vectors. The measure of similarity used is the cosine between two vectors, the normalised correlation coefficient. This similarity measure is used within a simple single-link clustering algorithm that clusters the two most similar context vectors at each stage up to a threshold similarity value.

Experiments and Results

As stated above, there were two primary objectives for this research. The first was to develop and evaluate a system that would address the cross-document coreference problem on a reasonably large scale. The second was to use this system to assess the potential utility of different contextual cues to the task. As ever, we can offer only partial answers to these questions but we believe that this is a promising start. We ran experiments to test the overall viability of the system and to appraise two premises: 1) That second-order co-occurrence would prove a more robust basis for representations than first-order co-occurrence. 2) That the names of other entities that co-occur with the name of interest might prove to carry more discriminatory information than non-name words. In addition, because named entities in the corpus were categorised into three classes (PER, ORG, LOC), it was also possible to vary the weighting of different entity types in vector representations. This measure allowed us to investigate the relative utility of different entity *types* to the disambiguation process.

To test these predictions, we have compared performance of the system using representations based on first-order and second-order (name) co-occurrence. Similarly, we have compared representations using name co-occurrence to those using non-name words. In experiments involving name-based representations we have also been able to experiment with different weighting schemes to vary the relative strengths of entity types in vector representations. All of our experiments used the same Wall Street Journal corpus of 2.5 million Words.

Evaluation

In order to measure system performance, we implemented an algorithm that was specifically designed for scoring coreference chains and previously applied to cross-document coreference - the B-Cubed algorithm (Bagga & Baldwin, 1998b). This algorithm produces recall, precision and F-measure scores for the system's performance by comparing the clusters formed with an answer key. The algorithm is shown below:

For each entity:

Recall (Ent-i) =

$$\frac{\text{No. of correct elements in the output cluster containing Entity i}}{\text{No. of elements in the truth cluster containing Entity i}}$$

Precision (Ent-i) =

$$\frac{\text{No. of correct elements in the output cluster containing Entity i}}{\text{No. of elements in the output cluster containing Entity i}}$$

The final precision and recall scores for each trial are calculated by averaging these individual scores across all the entities involved. In addition, the F-Measure is reported in the experiments below with precision and recall weighted equally. The above metric was used to measure the system's performance on each of the 30 namesets in each experimental condition.

It must be stated here, however, that the framework for evaluation did not use separate training and testing data as is standard practice in the evaluation of models such as this. The ubiquitous training-testing dichotomy is somewhat problematic in the case of cross-document coreference, principally because of the nature of the proper name domain. Specifically, it is not possible to determine in advance the proper names that a system will encounter, or the potential referents of each proper name. Different corpora will vary greatly in the proper names that they contain and, as new text is encountered, so too are new names and new referents for known names. Thus it is impossible for training to expose a system to every eventuality and the system must remain open ended and flexible. This entails a move away from the standard training-testing methodology for evaluation. The method described above is an automatic method that can be run on a corpus without intervention. Thus training consists of discovering the optimum settings for parameters such as weighting scheme, frequency limits for terms to be used as dimensions of vector space and similarity threshold for the clustering algorithm. Testing in the experiments reported here is then assessing the system's performance at disambiguating the thirty namesets. Possible improvements to this approach are discussed briefly in the final section.

Results

Table 1 above presents an overall comparison between the three main experimental conditions. Unfortunately, results have not yet been collected for first-order word co-occurrence. Results are quoted for the optimum similarity threshold. These results are the average scores across all 30 namesets at the similarity threshold shown. In all cases, very high frequency terms are not used as dimensions of vector space. In the word co-occurrence condition this constitutes the removal of stop words; for name co-occurrence this removes around 100 names that occur in an extremely high number of documents (generally the names of U.S. locations or states). In both cases, it was found that the inclusion of such terms was highly detrimental to performance. In both the name-based conditions above, all three entity types are weighted equally.

What should be evident from Table 1 is that the results obtained using word co-occurrence are poor compared to those obtained with name co-occurrence. This offers some support for claims of the importance of names in the cross-document coreference task but any conclusions must be tentative until results are available for first-order *word* co-occurrence. The high similarity threshold in this condition at which the best performance is achieved is also worthy of note. This shows that second-order *word* co-occurrence tends to result in fairly homogenous representations of context.

Surprisingly, there is not a marked difference between the first and second-order *name* co-occurrence conditions. However, the small benefit that second-order co-occurrence conveys was found consistently throughout experimentation. Added to this is the very low similarity threshold at which recall and precision scores are consistently highest in the first-order condition. This reflects the fact that there are typically very few shared elements in the vectors representing similar contexts. In the first-order model, representations are much more sparse and two occurrences can be judged similar on the basis of one shared co-occurrence. We feel that this situation is far from robust. The corpus we used was small and originated from a limited time period. This lead to a situation where the same entity often appeared in several documents that addressed the same or a closely related story, and thus a name often inhabited very similar immediate contexts. In a larger, more diverse corpus, where this degree of overlap is not as evident, it is highly likely that performance of the first-order model will degrade more rapidly than the second-order model.

Although this is a less than watertight case for second-order name co-occurrence, it is at the very least consistent with our claims that proper names may have a greater importance than other open class words and that second-order co-occurrence should be preferred over first-order co-occurrence. The constraints of this document prevent more in depth analysis of the various cases so the remainder of this section will focus solely on results from the second-order name co-occurrence condition.

Measure	Threshold	Recall %	Precision %	F-Measure %
First-Order Name Co-Occurrence	0.1	70.1	82.9	75.9
Second-Order Name Co-Occurrence	0.6	71.5	84.2	77.3
Second-Order Word Co-Occurrence	0.9	65.4	64.1	64.7

Table 1. System performance averaged across all namesets without selective entity weighting

Selective Weighting of Entity Types

Our experimentation sought to determine which features of the textual context surrounding a proper name are the most useful in disambiguating its referent. Having determined, at least, that other proper names provided considerable discriminatory information we have begun to investigate what kinds of names were useful. As explained above, the fact that the corpus is marked for named entities of three types has allowed us to vary the relative strengths of different entity types in the representations used. Using second order co-occurrence as the representational basis, it is also possible to vary the locus of this weighting. Specifically it is possible to weight first-order co-occurrences (by weighting an entire term vector) or second-order co-occurrences (by weighting individual elements in term vector). Naturally it is also possible to combine the two. Table 2 shows system performance using weighting to exclusively select one entity type at both levels (first and second-order). This exclusive weighting allowed us to assess the impact of names of a particular type in isolation and thus guide the search for the optimal weighting scheme. Rows in the table show the weighting scheme used and columns represent average F-scores across a group of namesets, either all 30 sets, or the 10 sets of a particular entity type. Thus, in first-order weighting, term vectors are constructed normally but only for the names of the indicated type that surround an occurrence of the name of interest. In second-order weighting, term vectors are created for all names that co-occur directly with the occurrence but the dimensions of vector space are restricted to names of a single entity type.

	Ave. F-scores for Namesets (%)			
	All	Loc	Org	Per
First-Order Term Vectors For:				
Locs Only	73.7	**73.3**	66.6	82.7
Orgs Only	74.6	69.7	**71.4**	84.2
Pers Only	**76.1**	65.8	69.0	**88.7**
Second-Order Vector Dimensions:				
Locs Only	76.1	74.7	71.0	83.9
Orgs Only	75.2	71.4	70.4	84.5
Pers Only	**78.4**	**78.6**	**71.4**	**86.5**

Table 2. System performance (F-Score) with selective weighting by entity type.

Although the significance of the results in Table 2 is not easy to interpret, we would suggest that there are a number of meaningful implications. The first half of the table suggests that the utility of different types of proper names depends to some extent on the type of name that is being disambiguated. In short, it appears that the most useful names in the immediate context of an occurrence are those of the same type as the target word. This is an intuitive result. In disambiguating an occurrence of 'Birmingham', the most useful contextual information would be the proper name 'Alabama' and not the name of someone that happens to live there. Conversely 'Pocahontas' might prove more useful to disambiguate an occurrence of 'John Smith' (the explorer) than the names of the places he explored. The second half of the table reveals which type of names are best to base representations on, specifically which names to use as the dimensions of vector space. The results show that restricting the dimensions of vector space to PER names alone elicits the best performance. The names of people typically appear in a more narrowly defined set of contexts than do those of locations and organisations. Thus the fact that two vectors have large scores in the dimension corresponding to '*Dan Winchester*' would typically suggest a greater degree of similarity than equivalent correspondences for '*University of Birmingham*' or '*Britain'*.

Guided by the above results and further experimentation with selective weighting it was possible to select the most appropriate weighting schemes for different types of entities. The system now alters the relative strength of different name types at both first and second order levels according to which type of name is being disambiguated. The weighting schemes favoured names of the same type as the target name at the first-order level and generally favoured PER names at the second-order level. Patterns also emerged in the optimum level similarity threshold to use for different entity types. The following table shows the system performance at its best.

Ave Scores for:	Threshold	R (%)	P (%)	F (%)
All Namesets	0.55	70.5	87.9	78.5
10 LOC sets	0.55	72.0	86.7	79.2
10 ORG sets	0.70	64.5	82.7	72.5
10 PER sets	0.35	91.2	88.8	90.3

Table 3. System Performance with optimal weighting

Table 3 above shows that performance is most promising for PER names and least impressive for ORG names. In general, we find these results promising. Superficially, scores of around 90% for precision and recall (admittedly only on PER names) would seem respectable for a prototype system. Precision scores are respectable for each group of namesets suggesting that the system is making few errors during clustering. However, the lower recall scores suggest that the system is often forming incomplete clusters.

Discussion

The aims of this research were twofold; to develop and evaluate a system that would address the cross-document coreference problem on a reasonably large scale and to use this system to assess the potential utility of different contextual cues to the task. We would claim partial success in both these objectives and are optimistic about the potential for future development of this work. It may be useful to relate the following discussion explicitly to the specific challenge questions of these symposia.

How does my model adjust to a new domain or to previously unknown material? As we have suggested above, the standard methodology of dividing a corpus into training and testing components is not immediately suitable for this particular task. However, it is obviously important to assess how well the model will generalise to novel text. We have identified two ways in which the model should be evaluated that correspond to the different ways in which such a system is likely to encounter new data. First, since this is an automatic method, it is possible to run the system on an entirely new corpus without intervention. The various parameters will be carried over from experimentation on the 'training' corpus. In this way, the training phase, such as it is, consists of our experimentation to determine the optimum settings for parameters such as the weighting scheme used, similarity threshold, frequency boundaries, etc. Testing will simply require that new namesets are defined and an answer key prepared for the new corpus. An initial, complete run of the system on the new corpus would perform the pre-processing and index-building steps. The system is then ready for trials on each of the new namesets.

Second, we will evaluate the response of the system to incremental changes in a familiar corpus. This is a slightly less straightforward proposition since it involves modifications to existing indexes and the classification of novel occurrences of a name. Importantly the indexes can be modified incrementally to reflect changes in the corpus so the computationally expensive indexing process can be avoided. Novel occurrences of a name will be incorporated by re-running the clustering algorithm. Although this involves a certain computational investment, it is preferable to the alternatives for two reasons: First the only way to avoid re-clustering is to store cumbersome vectors for each occurrence of each proper name in the corpus. In a respectably sized corpus this is a massive undertaking. Second, as the corpus changes, so too should the dimensions of vector space or the representations will eventually become incomplete. Any change in the dimensions of vector space would render stored representations inaccurate so that re-clustering would be necessary for all names. We are in the process of preparing a larger news corpus that will be used to implement and assess these two evaluation frameworks. Longer-term plans include investigating the system's applicability to other genres. News text is both rich in proper names and generally follows certain naming conventions. It is important to discover whether this general method can be applied to more unpredictable genres such as email.

How well does the model perform in a large-scale trial – by any metric? The initial results are promising for a prototype system. The system produced respectable precision and recall scores across 30 sets of ambiguous names. However, these results should be interpreted tentatively. There are a number of issues that require further investigation. First, the fact that CDC is a non-standard linguistic task means that there is not a well-established scoring metric. We have adopted the ubiquitous precision and recall paradigm and re-implemented the 'B-Cubed Algorithm' as previously used to score the cross-document coreference task (Bagga & Baldwin, 1998 a & b). However, we believe that there are certain issues surrounding the use of such a measure in this type of work. Specifically, it is not entirely clear what proportion of the scores that such a metric produces are attributable to system performance and what proportion are a product of the specific characteristics of the domain (in this case features of the nameset such as the size and the number of potential entities it contains). For example, it is a worrying artifact of the scoring metric that a system that puts all the documents in one cluster would obtain a Recall score of 100%. The corresponding precision score for this arbitrary performance would be entirely dependent on the characteristics of the nameset under investigation and thus, it is possible to obtain respectable scores from default performance. The dilemma is whether it is preferable to use an existing scoring metric because it is well-established and facilitates cross-study comparison, or whether to add to the proliferation of metrics that are designed around the specific features of a particular task. To be confident of the performance suggested by such scores we aim to develop a baseline measure against which performance can be measured. Unfortunately, there is no immediate candidate for such a measure.

In addition, the significance of this study is undermined by the size of the corpus used. While pseudo-ambiguity provides a useful way to overcome scarce data, it can only provide an approximation of the real problem. We are currently developing a larger corpus that will allow us to address this shortcoming and experiment with alternative evaluation paradigms.

What additional knowledge or theory does our model need to perform better? There are, of course, many ways in which the model could be improved, but the addition that might prove most significant would be the treatment of apposition. It has been observed that apposition often provides descriptive information about a named entity, particularly in the news genre. Coates-Stephens (1992) suggested that 80% of the proper names that he examined were accompanied by some form of description. Such information would prove highly valuable to the cross-document coreference task. We therefore need to develop techniques that can reliable identify and extract useful apposition without the need for extensive deep processing.

Conclusions. We believe that the research presented in this paper offers a general approach to the cross-document coreference problem that is viable and has good potential for refinement. This approach has allowed us to begin to investigate the potential role of different contextual features in the disambiguation of proper names. Our immediate plans are to implement the system on a larger corpus and use the evaluation frameworks described above to test performance on novel data.

Acknowledgements

Our thanks go to Sharp Laboratories of Europe who provide both funding and support to Dan Winchester through an EPSRC Case studentship.

References

ACE 1999. *Entity Detection and Tracking – Phase 1,* *ACE Pilot Study Task Definition.* Downloadable via anonymous ftp from: ftp://jaguar.ncsl.nist.gov/ace/phase1/edt_phase1_v2.2.doc

Bagga, A., and Baldwin, B. 1998a. How Much Processing is Required for Cross-Document Coreference? In *Proceedings of the ACL workshop on Coreference and Its Applications*, Maryland. van Deemter, Kees and Rodger Kibble.

Bagga, A., and Baldwin, B. 1998b. Algorithms for Scoring Coreference Chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, 563-566.

Coates-Stephens, S. 1992. *The Analysis and Acquisition of Proper Names for Robust Text Understanding*. Ph.D. Diss., City University, London.

Mikheev, A., Moens, M and Grover, C. 1999. Named Entity Recognition without Gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99),* 1-8.

Ravin, Y., and Kazi, Z. 1998. Is Hillary Rodham Clinton the President? Disambiguating Names across Documents. In *Proceedings of the ACL workshop on Coreference and Its Applications*. Maryland. van Deemter, Kees and Rodger Kibble.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, London.

Schutze, H. 1998. *Automatic Word Sense Discrimination*. Computational Linguistics, 24/1: 96-123.